ILRR: Inference-Time Steering Method for Masked Diffusion Language Models

Eden Avrahami[1], **Eliya Nachmani**[2],
[1]Blavatnik School of Computer Science, Tel Aviv University
[2]School of Electrical and Computer Engineering, Ben-Gurion University of the Negev
ea1@mail.tau.ac.il, eliyanac@bgu.ac.il

Abstract

Discrete Diffusion Language Models (DLMs) offer a promising non-autoregressive alternative for text generation, yet effective mechanisms for inference-time control remain relatively underexplored. Existing approaches include sampling-level guidance or trajectory optimization mechanisms. In this work, we study the paradigm of reference-based latent steering for DLMs. We introduce Iterative Latent Representation Refinement (ILRR), an efficient framework for steering DLMs using a reference text as a high-level semantic blueprint. ILRR extracts and injects reference-derived semantic signals into the evolving activations of the generated sequence, enabling tunable transfer of coarse properties such as sentiment. We further introduce Spatially Modulated Steering, an extension that enables long-form generation to be guided by shorter references by regulating intensity across the sequence. Empirically, we demonstrate that ILRR achieves effective control on LLaDA and MDLM architectures with low computational overhead, requiring only one additional parallel forward pass per denoising step. Under comparable compute budgets, ILRR improves attribute accuracy over baselines by 10% to 60% points. Our results suggest that the iterative, global denoising process makes DLMs a natural substrate for effective sequence-wide activation-level control.

1 Introduction

Discrete diffusion language models (DLMs) have recently emerged as a promising non-autoregressive framework for text generation (Austin et al., 2021; Sahoo et al., 2024). Recent scaled models such as Dream (Ye et al., 2025b) and LLaDA (Nie et al., 2025) demonstrate strong performance, even on complex tasks such as mathematical reasoning and coding. Unlike autoregressive (AR) models, which generate text sequentially (Vaswani et al., 2017; Bengio et al., 2003; Radford et al., 2018), DLMs generate through a global iterative procedure, reconstructing tokens by sampling from a learned reverse denoising process.

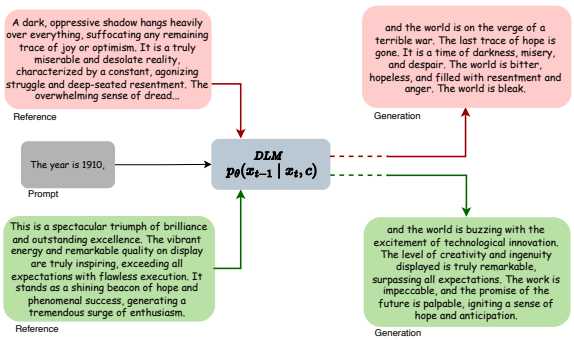


Figure 1: References and corresponding generations for positive and negative sentiment. ILRR guides generations toward the coarse semantics of the reference.

Despite this progress, inference-time control for DLMs remains relatively underexplored. Inference-time steering is particularly appealing as it allows for flexible modulation of model behaviors without the cost of parameter fine-tuning. Existing DLM steering methods primarily operate at the sampling or trajectory level, including classifier guidance, classifier-free mechanisms, remasking policies and inference-time trajectory optimization (Singhal et al., 2025; Dang et al., 2026; Schiff et al., 2025; Wang et al., 2026; Jazbec et al., 2025). While effective, these approaches often require external classifiers, multiple candidate sequences or iterative resampling, limiting their flexibility or increasing inference cost. A complementary direction is representation-level control, where model behavior is steered by modifying internal activations. In the AR domain, this approach has proven effective, offering practical advantages such as tunable control strength, context efficiency, robustness to prompt brittleness, and stronger enforcement of behaviors weakly expressed under standard prompting or decoding (Stolfo et al., 2025; Liu et al., 2024; Rim-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 495–512
November 6-10, 2026 ©2026 Association for Computational Linguistics

sky et al., 2024; Wang et al., 2025a; Park et al., 2024). This area remains relatively less explored for DLMs, which create a unique opportunity by exposing global, sequence-wide representations throughout the iterative denoising process. Alongside this, reference-based guidance has proven effective in continuous image diffusion, where the iterative and global nature of the denoising process enables efficient plug-and-play control using a single reference image for coarse semantic guidance (Tumanyan et al., 2022; Choi et al., 2021). Together, these advances create a natural opportunity to explore whether reference-based activation guidance can achieve effective control for DLMs.

In this work, we propose **Iterative Latent Representation Refinement (ILRR)**, an efficient and adaptable framework for latent steering in DLMs. Inspired by the Iterative Latent Variable Refinement (ILVR) technique in continuous image diffusion (Choi et al., 2021), ILRR uses a reference text as a high-level semantic blueprint that guides the generation. It iteratively extracts steering signals from a noise-corrupted reference and injects them into selected hidden states of the generated sequence. This nudges the denoising trajectory toward the coarse properties encoded by the reference while allowing the model to determine the concrete wording, syntax and prompt-specific content. As a plug-and-play method, ILRR enables flexible control over semantic attributes without task-specific classifiers or additional training, requiring only one additional parallel forward pass per denoising step. We evaluate ILRR on standard controlled generation benchmarks, including toxicity and sentiment steering, using LLaDA and MDLM as base models, and show that ILRR consistently outperforms competitive inference-time baselines under comparable compute budgets. We further test ILRR's robustness on detoxification and topic control, both showing effective steering. Our results indicate that reference-based latent steering offers an effective approach for DLMs, leveraging the iterative global nature of the generation process. **Our main contributions are as follows:** (i) We introduce ILRR, a learning-free framework for latent steering in DLMs, combining activation-level control with plug-and-play reference guidance, with tunable parameters controlling steering intensity. (ii) We further propose the spatially modulated steering extension, which allows effective guidance when the reference sequence is significantly shorter than the target output. (iii) We demonstrate that ILRR achieves strong controllability with low computational overhead, improving steering accuracy by 10% to 60% points over competitive baselines.

2 Related Work

Discrete diffusion language models have emerged as a robust non-AR alternative for text generation, employing architectures that typically utilize iterative denoising or unmasking processes to reconstruct masked sequences (Austin et al., 2021; Sahoo et al., 2024; Nie et al., 2025; Ye et al., 2025b; Lou et al., 2024; Shi et al., 2024; Yu et al., 2025; von Rütte et al., 2026; Shariatian et al., 2026; Zhou et al., 2026; Arriola et al., 2025; von Rütte et al., 2025; Wang et al., 2026; Liu et al., 2025; Kollovieh et al., 2026; He et al., 2023; Gong et al., 2025).

Controllable generation is well-established in AR modeling. Methods include utilization of gradients or auxiliary models, discriminators or decoding-time experts (Dathathri et al., 2020; Yang and Klein, 2021; Krause et al., 2021; Liu et al., 2021; Lovelace et al., 2024).

Activation steering is another central paradigm, which modifies internal activations to control model's generation, often using steering directions derived from contrastive data or interpretable features. Such methods typically employ an update rule of the form $h \leftarrow h + \alpha v$, modifying the hidden states h during generation using a steering vector v scaled by a factor α (Turner et al., 2024; Zou et al., 2025; Rimsky et al., 2024; Arditi et al., 2024; Stolfo et al., 2025; Liu et al., 2024; Wang et al., 2025a; Park et al., 2024; Wang et al., 2025b; Li et al., 2024a; Lee et al., 2025; Peters et al., 2018; Sharma and Trivedi, 2026).

Controllable generation methods for DLMs include reward-guided and particle sampling, search algorithms, classifier-based and classifier-free guidance, learned policies, differentiable constraints or trajectory optimization mechanisms (Singhal et al., 2025; Dang et al., 2026; Ye et al., 2025a; Zhang et al., 2026; Schiff et al., 2025; Jazbec et al., 2025; Cardei et al., 2026; Chen et al., 2025; Tejaswi et al., 2026; Peng et al., 2026; Zhao et al., 2026; Li et al., 2024b; Phunyaphibarn and Sung, 2026; Padole et al., 2025).

Reference-based diffusion guidance approaches have demonstrated effectiveness for controllable generation in the image domain, utilizing reference-extracted signals to guide the generation process (Choi et al., 2021; Tumanyan

et al., 2022; Wang and Sun, 2025).

3 Background

3.1 Diffusion Generative Models

Diffusion models are a class of generative models that learn to reverse a gradual corruption process. Generation begins with a sample from a high-entropy prior distribution x_T (e.g., pure noise) and iteratively refines it over timesteps $t = T, \dots, 1$ to produce a clean sample x_0. Formally, this framework consists of two main processes: **(i) Forward process** $q(x_t|x_0)$, which gradually corrupts clean data x_0 into noise x_T according to a fixed noise schedule. **(ii) Reverse process** $p_\theta(x_{t-1}|x_t)$, a model trained to reverse this corruption process, by estimating the clean data or the noise component at each step. While this structure is shared, the mathematical definition of these processes differs between the continuous and discrete domains.

3.1.1 Continuous Diffusion Models

In continuous domains (e.g., images), the forward process $q(x_t|x_0)$ typically adds Gaussian noise to the data. For a clean data point x_0, the noisy state x_t is defined as $x_t = \sqrt{\bar{\beta}_t}x_0 + \sqrt{1-\bar{\beta}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$ where $\bar{\beta}_t$ is the noise schedule. The reverse process $p_\theta(x_{t-1}|x_t)$ is a network trained to remove this noise by predicting ϵ to gradually recover the original data (Sohl-Dickstein et al., 2015; Ho et al., 2020).

3.1.2 Discrete Diffusion Language Models

DLMs (Austin et al., 2021) adapt the framework to the discrete nature of language. We consider a sequence $x_0 = (x_0^1, x_0^2, \dots, x_0^N)$ where each token x_0^i belongs to a vocabulary $\mathcal{V}$, and introduce a special mask token $[M]$.

Forward Process. The forward process q gradually corrupts the clean sequence x_0 by substituting discrete tokens with the mask token $[M]$. The marginal distribution $q(x_t|x_0)$ allows sampling any intermediate state x_t directly:

$$q(x_t|x_0) = Cat(x_t; x_0, \bar{\alpha}_t) \tag{1}$$

where $Cat(\cdot)$ is the categorical distribution over the vocabulary $\mathcal{V}$ with the noise schedule $\bar{\alpha}_t$, and each token x_t^i is either the original x_0^i or $[M]$. As $t \to T$, the sequence becomes fully masked.

Reverse Process. The reverse process $p_\theta(x_{t-1}|x_t)$ reconstructs the original tokens from the masked tokens. DLMs typically employ a denoising network $p_\theta(x_0|x_t)$, trained to predict x_0 from x_t. The reverse transition is then computed using the posterior of the forward process:

$$p_\theta(x_{t-1}|x_t) \propto q(x_{t-1}|x_t, x_0 \sim p_\theta(x_0|x_t)) \tag{2}$$

Practically, the model predicts tokens for the masked positions and then re-samples a partially masked sequence x_{t-1}.

Network Structure. The network $p_\theta(x_0|x_t)$ is typically parameterized as a deep neural network with D layers (blocks), which we denote by $L_{p_\theta} = (l_1, \dots, l_D)$. Given an input sequence $x = (x^1, \dots, x^N)$ of length N, let $(h_x^{(1)}, \dots, h_x^{(D)})$ denote the sequence of hidden states (the residual stream) produced by the network, where $h_x^{(i)} = (h_{x^1}^{(i)}, \dots, h_{x^N}^{(i)}), \quad h_{x^j}^{(i)} \in \mathbb{R}^d$ such that $h_{x^j}^{(i)}$ represents the hidden vector for token x^j at layer l_i, and d is the model's hidden dimension. Finally, a prediction head $f_{head}(\cdot)$ maps the final hidden states to vocabulary logits: $f_{head}\big(h_x^{(D)}\big) \in \mathbb{R}^{N\times|\mathcal{V}|}$. The denoising distribution over the vocabulary is then given by $p_\theta(x_0|x_t) = softmax\big(f_{head}(h_x^{(D)})\big)$.

Conditioning Model. The denoising model may additionally be conditioned on a prefix prompt c, in which case we write the denoising network as $p_\theta(x_0|x_t, c)$ and the corresponding reverse transition as $p_\theta(x_{t-1}|x_t, c)$.

3.2 Iterative Latent Variable Refinement

ILVR (Choi et al., 2021) is a guidance method originally developed for continuous image diffusion. It conditions generation by aligning the low-frequency components of the generated image with those of a reference image y. Let $\phi(\cdot)$ be a linear low-pass filter. In each reverse step t, ILVR generates a candidate x'_{t-1} from the model and mixes it with a corrupted version of the reference $y_{t-1} \sim q(y_{t-1}|y)$ (Eq.8 from Choi et al. (2021)):

$$x_{t-1} = \phi(y_{t-1}) + (I - \phi)(x'_{t-1}) \tag{3}$$

By enforcing consistency in the low-frequency domain defined by ϕ, ILVR locks the coarse attributes of the generation to the reference y, while the high-frequency details are derived from the model's proposal. This allows for diverse generations that remain semantically aligned with the reference.

4 Method

Inspired by the principle of ILVR, given a trained DLM $p_\theta(x_{t-1}|x_t, c)$, we aim to generate a text $x_0 \in \mathcal{V}^N$ that shares its high-level semantics with

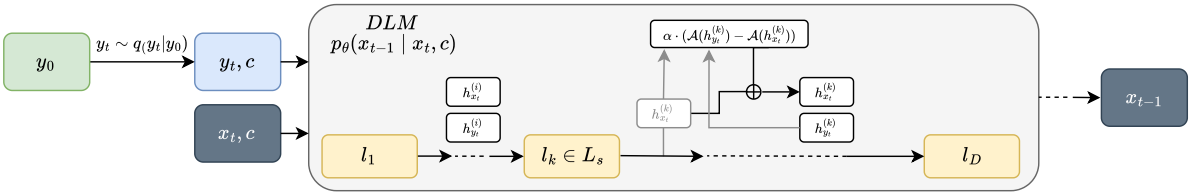


Figure 2: Overview of ILRR's steering process within the latent space. The noised reference y_t is processed parallel to x_t. Semantic signals are extracted via $\mathcal{A}$ and injected into the generation's hidden states at target layers L_s.

a given reference text $y_0 \in \mathcal{V}^N$. ILRR achieves this by iteratively refining the latent representations of the generation sequence with coarsened semantic signals derived from the reference sequence. Given a total of T denoising steps $1, \ldots, T$, refinement is applied in a desired subset of steps $T_s \subseteq \{1, \ldots, T\}$, to a targeted subset of hidden states (residual-stream activations) produced from layers $L_s \subseteq L_{p_\theta}$, empirically found to best encode high-level semantic attributes (Section 6).

4.1 Iterative Latent Representation Refinement

The ILRR procedure at each timestep $t \in T_s$ is defined as follows. First, to ensure that the representations are distributionally aligned, we corrupt the reference sequence y_0 to the same corruption level as the current generation state x_t by sampling $y_t \sim q(y_t|y_0)$ (Equation (1)). We then perform two parallel forward passes through the denoising model: $p_\theta(\cdot|y_t, c)$ and $p_\theta(x_{t-1}|x_t, c)$. As both sequences propagate through the model, we intervene after each specified target layer $l_k \in L_s$: Let $h_{x_t}^{(k)}$ and $h_{y_t}^{(k)}$ denote the computed activations for the generation x_t and the noised reference y_t, respectively, in layer l_k. Given a dimension-preserving semantic-extractor operation $\mathcal{A}(\cdot)$ and a steering scale α, we refine the generation's activations via the update rule:

$$h_{x_t}^{(k)} \leftarrow h_{x_t}^{(k)} + \alpha \left(\mathcal{A}(h_{y_t}^{(k)}) - \mathcal{A}(h_{x_t}^{(k)}) \right) \quad (4)$$

This update rule is grounded in the established activation steering methodology (Section 2), with our steering vectors derived dynamically throughout the denoising process. These modified activations then continue to propagate through the network, shifting all subsequent hidden states and ultimately biasing the prediction of the next state x_{t-1}. Importantly, while the forward pass on x_t is used to sample the next state x_{t-1}, the forward pass on y_t is performed solely to extract its hidden representations by the model, while the reference y_0 remains static and is re-corrupted at each step. By applying this process iteratively across the denoising steps, ILRR gradually directs the latent representations of the generation toward those of the reference in the high-level semantic space defined by $\mathcal{A}$, while the scale α controls the strength of this guidance. The complete inference procedure is formalized in Algorithm 1, with the method diagram illustrated in Figure 2 and generation examples shown in Figure 1.

Algorithm 1 Iterative Latent Representation Refinement

Require: prefix prompt c, trained DLM $p_\theta(x_{t-1}|x_t, c)$, timesteps T, reference text $y_0 \in \mathcal{V}^N$, steering scale α, target steering layers $L_s \subseteq L_{p_\theta}$, steering steps T_s
Ensure: generated text $x_0 \in \mathcal{V}^N$

$x_T \leftarrow [M]^N$ ▷ Initialize fully masked
for $t = T, \ldots, 1$ **do**
  $y_t \sim q(y_t|y_0)$
  for k from 1 to D **do**
    compute $h_{x_t}^{(k)}$ from $h_{x_t}^{(k-1)}$
    compute $h_{y_t}^{(k)}$ from $h_{y_t}^{(k-1)}$
    if $l_k \in L_s$ **and** $t \in T_s$ **then**
      $h_{x_t}^{(k)} \leftarrow h_{x_t}^{(k)} + \alpha \cdot (\mathcal{A}(h_{y_t}^{(k)}) - \mathcal{A}(h_{x_t}^{(k)}))$
    end if
  end for
  compute x_0 from $softmax(f_{head}(h_{x_t}^{(D)}))$
  compute $x_{t-1} \sim q(x_{t-1}|x_t, x_0)$
end for
Return x_0

4.2 Extracting High-Level Semantics

Building upon the understanding that internal representations in language models capture contextualized, interpreted semantic information (Peters et al., 2018; Tenney et al., 2019; Park et al., 2024; Zou et al., 2025; Turner et al., 2024), we adopt 1D average pooling as the semantic-extractor operation $\mathcal{A}$. Our motivation is that, if a semantic attribute is encoded as a shared component in the hidden space,

aggregating representations across the sequence should reinforce this common component while attenuating position-specific and lexical variation. We apply this operation as a sliding window along the sequence dimension. Formally, for a sequence of activation vectors $h_x^{(l)} = (h_{x_1}^{(l)}, \ldots, h_{x_N}^{(l)})$, and a kernel size k, we compute the pooled representation at position i by averaging the activations within a local window:

$$\mathcal{A}(h_x^{(l)})_i = a_i^{(l)} = \frac{1}{k} \sum_{j=i-\lfloor k/2 \rfloor}^{i+\lfloor k/2 \rfloor} h_{x_j}^{(l)} \qquad (5)$$

The resulting sequence $(a_1^{(l)}, \ldots, a_N^{(l)})$ aggregates adjacent token activations into a smoothened regional semantic direction. By computing these pooled activations for both the reference and the generation, we can align their underlying semantic directions.

4.3 Spatially Modulated Steering

In certain domains, a reasonable use case is to use a short reference to steer the generation of a longer sequence ($N_{x_0} > N_{y_0}$). Extending our framework to this setting requires adjusting the mechanism in two key aspects: (i) Dimensionality Alignment, to enable the computation of valid steering directions between sequences of different lengths; and (ii) Steering Adaptability, to handle the challenge of distributing limited reference signals across a longer sequence in a way that is not overly aggressive (avoiding excessive repetition of identical steering directions), while remaining effective (preserving semantic information without dilution). To address this, we propose **Spatially Modulated Steering**, an extension that overcomes these challenges through the following mechanisms.

Interpolated Steering. Unlike the standard procedure, we operate on sequences of differing lengths. After each layer $l_k \in L_s$, we first apply the semantic extractor $\mathcal{A}$ to the hidden states of both sequences to capture their respective semantic signals. To align their dimensions for the steering update, we downsample the generation's pooled activations using adaptive average pooling, which compresses the longer sequence into a fixed target length equal to that of the reference, N_{y_0}. This operation partitions the generation sequence into uniformly spaced intervals and computes the mean of the hidden states within each interval, such that the i-th pooled vector represents the average hidden state over the i-th partition. This yields a length-aligned representation while preserving dominant semantic directions through aggregation along the sequence dimension. The steering directions are then formed by upsampling the difference between these aligned representations back to length N_{x_0} via linear interpolation. This interpolation estimates higher-resolution values by taking weighted averages of neighboring positions, effectively spreading the reference's semantic signal smoothly across the longer generation window. Formally, let $\mathcal{I}_\downarrow : \mathbb{R}^{N_{x_0} \times d} \to \mathbb{R}^{N_{y_0} \times d}$ denote the adaptive pooling operation and $\mathcal{I}_\uparrow : \mathbb{R}^{N_{y_0} \times d} \to \mathbb{R}^{N_{x_0} \times d}$ denote the linear interpolation operation. The spatially aligned steering directions $\hat{\Delta}^{(k)}$ are computed as:

$$\hat{\Delta}^{(k)} = \mathcal{I}_\uparrow \left(\mathcal{A}(h_{y_t}^{(k)}) - \mathcal{I}_\downarrow(\mathcal{A}(h_{x_t}^{(k)})) \right) \qquad (6)$$

This operation effectively stretches the dense semantic blueprint of the short reference to cover the longer generation window, ensuring that the guiding signal is continuously distributed.

Modulated Steering Intensity. To prevent dense reference signals from being excessively injected across a broader sequence, we modulate the steering strength spatially along the generation. Specifically, we compute a wave-shaped modulation vector $\mathbf{w} \in \mathbb{R}^{N_{x_0}}$ that smoothly varies the steering intensity across positions. The waveform vector (e.g., cosine) is scaled such that its peaks apply the full steering strength α, while its troughs apply minimal steering. Formally, employing a cosine wave pattern with frequency f, the modulation weight $\mathbf{w}_i$ at position i is defined as:

$$\mathbf{w}_i = \frac{\alpha}{2} \left[1 + \cos \left(\frac{2\pi f \cdot i}{N_{x_0}} \right) \right] \qquad (7)$$

where the term $\frac{i}{N_{x_0}}$ spatially normalizes the position, ensuring that the wave pattern is distributed proportionally across the sequence regardless of length, while the outer operations scale the output to the range $[0, \alpha]$. We note that the waveform structure is a flexible hyperparameter that can be adapted. For instance, one may adjust the wave amplitude or frequency to moderate the intensity variation for shorter sequences. By scaling relative to the sequence length, we can ensure that each region of the generated sequence receives a proportional portion of the reference signals. This spatial modulation balances coverage and strength, allowing reference semantics to influence all regions of

the generation while avoiding potential distortions caused by uniformly distributing the limited reference signals across a longer sequence.

The Modified Update Rule: Finally, using Equation (6) and Equation (7), we modify the update rule in line 8 of Algorithm 1 as follows:

$$h_{x_t}^{(k)} \leftarrow h_{x_t}^{(k)} + \mathbf{w} \cdot \hat{\Delta}^{(k)} \tag{8}$$

This ensures that the reference signal is both spatially aligned via adaptive downsampling and dynamically distributed via the modulation vector $\mathbf{w}$.

4.4 Controllable Steering Dynamics

Our method introduces two key mechanisms that allow for tunable control over the steering behavior. **(i) Steering Scale α.** The scalar α determines the magnitude of semantic routing (Equation (4)), acting as a coefficient for the strength of the guidance. While a single global parameter is sufficient for general steering, our framework supports fine-grained control by allowing for layer-specific scaling. We can define a configuration vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{|L_s|})$ where each α_i corresponds to a specific injection layer $l_k \in L_s$. **(ii) Timesteps T_s.** We can further refine the steering by restricting the ILRR intervention to a specific subset of timesteps $T_s \subseteq \{1, \ldots, T\}$, allowing us to apply semantic steering only during specific stages of the generation process (e.g., early semantic formation), leaving the remaining steps unconstrained.

5 Experiments

5.1 Main Experiments

To evaluate ILRR for controllable generation, we examine its ability to steer DLMs toward semantic attributes, using references that exhibit the target attribute. We compare our method against state-of-the-art inference time particle-sampling-based baselines, including PG-DLM, FK-Steering and best-of-n (Dang et al., 2026; Singhal et al., 2025).

5.1.1 Experimental Setup

Datasets. Following prior work (Dang et al., 2026; Singhal et al., 2025; Han et al., 2023), for each task we evaluate on the standard suite of 15 prefix prompts, generating 20 continuations per prompt. To ensure robust steering assessment, for each assessed attribute we utilize 5 distinct reference sequences per prompt, running 4 generations per reference-prompt pair. We repeat this setup across 3 random seeds.

Model	Method	Toxicity (↑)	Sentiment (↑)
MDLM	best-of-n	1.9 ± 0.4	36.7 ± 3.7
	FK ($\phi = 4$)	0.8 ± 0.2	10.0 ± 1.3
	FK ($\phi = 1$)	3.8 ± 1.0	37.4 ± 1.2
	PG-DLM	1.4 ± 0.7	23.8 ± 2.2
	ILRR ($\alpha = 0.7$)	4.7 ± 0.6	43.4 ± 1.3
	ILRR ($\alpha = 0.8$)	9.4 ± 1.2	51.3 ± 4.1
	ILRR ($\alpha = 1.0$)	$\mathbf{15.9} \pm 2.0$	$\mathbf{65.2} \pm 2.5$
LLaDA	best-of-n	2.4 ± 0.2	48.2 ± 2.9
	FK	9.0 ± 1.5	69.4 ± 1.2
	PG-DLM	8.3 ± 1.8	66.6 ± 1.0
	ILRR ($\alpha_1, \alpha_2 = 0.7, 0.75$)	45.4 ± 2.5	97.7 ± 0.3
	ILRR ($\alpha_1, \alpha_2 = 0.8, 0.85$)	60.7 ± 0.7	98.8 ± 0.7
	ILRR ($\alpha_1, \alpha_2 = 0.95, 1.0$)	$\mathbf{71.2} \pm 1.2$	$\mathbf{100.0} \pm 0.0$

Table 1: Short generation results (length 50). Toxicity and positive sentiment accuracy (mean $\pm$ std) on MDLM and LLaDA across 3 random seeds.

Metrics. We evaluate steering effectiveness on toxicity and sentiment control. Following Dang et al. (2026), we employ the following external classifiers: **(i) Toxicity -** a RoBERTa-based toxicity classifier (Logacheva et al., 2022) trained to detect harmful content (e.g., hate speech, insults). A generation is considered successful if it is classified as *toxic*. While usually undesirable, toxicity is typically rare in model generations, making this task a rigorous benchmark for steering toward low-probability concepts. **(ii) Sentiment -** a RoBERTa-based sentiment classifier (Barbieri et al., 2020) fine-tuned on the TweetEval benchmark. A generation is considered successful if it is classified as *positive*. For both tasks we report accuracy, defined as the percentage of generated sequences successfully classified as possessing the target attribute.

5.1.2 Implementation Details

We apply ILRR to base DLMs, LLaDA-8B-Base (Nie et al., 2025) and MDLM (Sahoo et al., 2024), integrating our method into their native generation loops. We conduct two sets of experiments: (i) 50-length generation on LLaDA and MDLM, with length-matched references, used to evaluate our standard mechanism (Section 4.1), and (ii) 512-length generation on MDLM, with approximately 60-token references, used to evaluate spatially modulated steering (Section 4.3). In all experiments, we operate under a compute budget measured by the number of function evaluations (*NFE*), as defined in Dang et al. (2026), under which our method corresponds to $NFE = 2 \cdot T$. For convenience, we normalize the *NFE* value by the number of denoising steps T. We compare our approach against the

Method	Toxicity (↑)	Sentiment (↑)
best-of-n	1.0	23.0
FK ($\phi = 4$)	0.0	7.3
FK ($\phi = 1$)	3.0	26.0
PG-DLM	1.7	17.3
ILRR ($\alpha = 0.7$)	3.8 ± 0.5	43.7 ± 2.5
ILRR ($\alpha = 0.8$)	5.7 ± 2.0	48.2 ± 1.9
ILRR ($\alpha = 1.0$)	**12.1** ± 2.0	**58.4** ± 3.4

Table 2: Long generation results (length 512). Toxicity and positive sentiment accuracy on MDLM using spatially modulated steering across 3 random seeds.

Method	Toxicity (%) (↓)
Base Model	40.9
best-of-n ($n = 2$)	24.9
best-of-n ($n = 4$)	13.2
ILRR ($\alpha_1, \alpha_2 = 0.7, 0.75$)	10.7
ILRR ($\alpha_1, \alpha_2 = 0.8, 0.85$)	**8.0**

(a) Detoxification results on the most toxic RealToxicityPrompts examples.

Relevance ≥ 4 (%) (↑)		
Topic / Scale	$(0.7, 0.75)$	$(0.8, 0.85)$
Space	94.3	**97.5**
Religion	88.8	**94.0**
Science	84.3	**89.0**

(b) ILRR topic control results. Topic relevance is evaluated using GPT-4o-mini as-a-judge.

Table 3: Generalization averaged results (LLaDA) using small pools of globally reused references.

baseline performance reported in Dang et al. (2026) at the comparable compute budget of normalized $NFE = 4$. The full implementation and configuration details including our intervention layers, denoising and steering steps, scales, modulation parameters, datasets and reference curation are provided in Appendix A.

5.1.3 Main Results

Table 1 reports steering accuracy for 50-token generation on LLaDA and MDLM, showing ILRR consistently outperforms all inference-time baselines in the constrained compute budget. On LLaDA, ILRR reaches 71.2% toxicity and 100.0% sentiment accuracy, compared to 9.0% and 69.4% for the strongest baselines. On MDLM, ILRR improves toxicity from 3.8% to 15.9% and sentiment from 37.4% to 65.2%. Table 2 details the results for 512-token generation on MDLM using spatially modulated steering, showing that ILRR effectively improves toxicity accuracy from 3.0% to 12.1% and sentiment accuracy from 26.0% to 58.4%, an improvement of 32% points.

5.2 Generalization and Broad Applicability

To evaluate ILRR's robustness, we conduct additional experiments on LLaDA, testing (i) whether the same hyperparameter configuration (Appendix A) transfers to additional control objectives of detoxification and topic control, and (ii) whether small pools of reused references can steer generations of diverse prompts, reflecting a practical setting in which references are curated once for a desired attribute or domain and reused across different conditioned generations. For detoxification, we use the 791 most toxic prompts from RealToxicityPrompts (Gehman et al., 2020) and steer toward non-toxic continuations using only 20 global safe references. Each prompt is paired with its top-3 cosine-similar references, and we generate 4 40-length continuations per pair with toxicity evaluated using Logacheva et al. (2022). For topic control, we follow Dathathri et al. (2020) and steer their standard suite of 20 topic control prompts toward Space, Religion, and Science using 5 global references per topic, generating 4 50-length continuations per prompt-reference pair. Since dedicated classifiers are not naturally available for such open-ended topics, we evaluate topic relevance using an LLM-as-a-judge (GPT-4o-mini), prompted to assign a 1-5 relevance score. As shown in Table 3, ILRR substantially reduces toxicity and achieves high topic relevance, suggesting that the method can support effective zero-shot steering - small pools of references can be reused to guide diverse prompts, without task-specific tuning.

Overall, the results indicate that iterative latent intervention is an effective mechanism for steering DLMs. ILRR transfers reference-derived guidance across all tasks and length settings, and enables efficient steering under a relatively low computational budget compared to the sampling-based baselines. We attribute this effectiveness to the rich sequence-wide semantic information encoded in DLM hidden states; by gradually aligning these representations with reference-derived signals, the generation trajectory is biased toward the desired semantic direction.

We additionally provide the score distributions given by the external classifiers in Appendix B, and qualitative examples in Appendix D.

6 Analysis and Ablation

To better understand the effect of each component of our framework, we analyze how key parameters impact the steering effectiveness of ILRR while keeping all other parameters fixed. Further analysis details are provided in Appendix C.

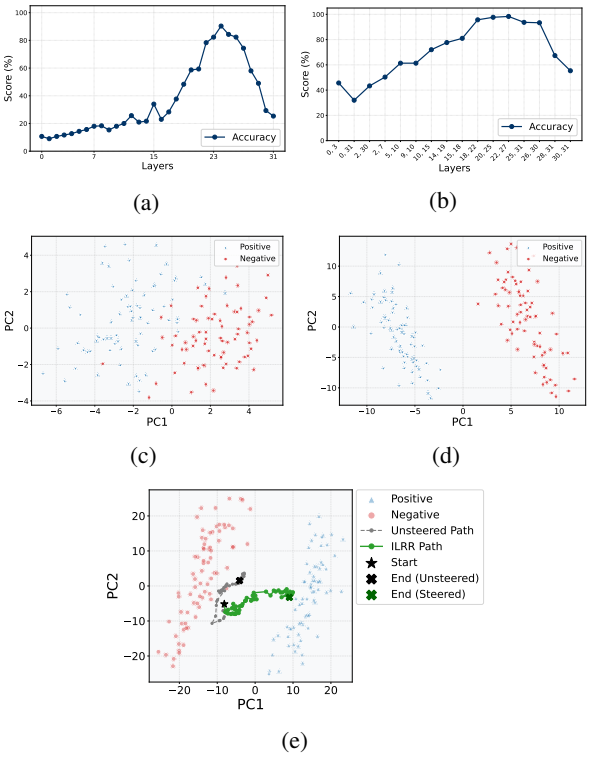


Figure 3: Layers and latent geometry. (a) Accuracy across single model layers. (b) Accuracy across pairs of model layers. (c, d) PCA projections of latent representations for positive and negative sequences at layers 6 and 20, respectively. (e) Tracking the steered vs. unsteered generation trajectories via PCA at layer 25.

Layer selection and latent geometry (Appendix C.1). Figures 3a and 3b show that steering is most effective in mid-to-late layers, while early-layer interventions are consistently weaker. Figures 3c to 3e provide a geometric explanation for this trend: representations are more entangled in early layers but become separable in mid-to-late layers, where ILRR can redirect the denoising trajectory toward the target semantic region.

Tunability (Appendix C.2). Figure 4a shows that steering intensity increases consistently with α, while 4-gram overlap with the reference remains low even at the highest scale, indicating that stronger control arises from semantic alignment rather than heavy verbatim reuse of the reference. Figure 4b shows that ILRR is effective in various denoising stages, with early intervention outperforming late, indicating that coarse semantic control is more influential early in generation.

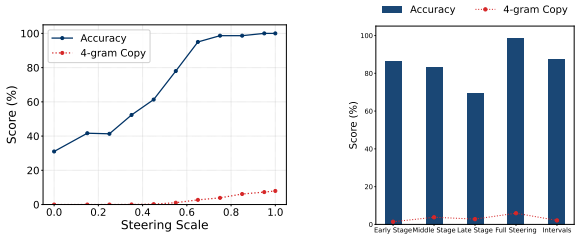


Figure 4: Tunability analysis. (a) Accuracy compared to 4-gram overlap with the reference text across steering scales α. (b) Accuracy compared to 4-gram overlap with the reference text across steering step sets T_s.

Reference type	Acc (%)(↑)	Flu. (↑)	Dist-1/2/3 (↑)	PPL (↓)
Original	**97.3**	**4.2**	.42 / .74 / .84	18.3
Sparse (25% word drop)	86.7	3.0	.31 / .56 / .66	12.5
Noise (50% random words)	69.3	2.6	.33 / .60 / .73	13.3
Sparse (50% word drop)	45.3	1.8	.14 / .26 / .32	5.4
Chunk-shuffled (10-words)	96.0	3.7	.40 / .72 / .85	17.0
Chunk-shuffled (5-words)	86.7	3.5	.37 / .66 / .79	14.9
Chunk-shuffled (2-words)	73.3	2.8	.32 / .61 / .75	13.8
Fully Shuffled	61.3	2.2	.29 / .57 / .73	14.4
Highly repetitive	84.0	2.1	.13 / .23 / .29	**3.5**
Midway sentiment shift	46.7	3.9	**.44 / .76 / .86**	20.7
Grammatically correct gibberish	37.3	3.4	.36 / .57 / .68	6.8

Table 4: The impact of reference corruption, structural shuffling and semantic degradation on the generation. Flu. denotes GPT-4o fluency score on a 1–5 scale.

Reference quality (Appendix C.3). Table 4 shows that effective references require coherent, consistent semantics and sufficient lexical diversity, while corrupted, shuffled, inconsistent, or repetitive references degrade control and quality.

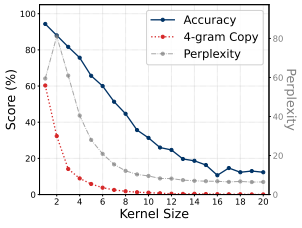
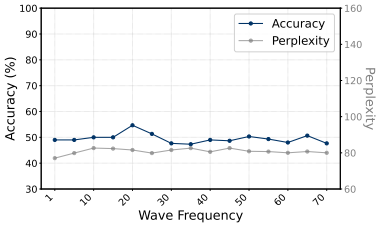


Figure 5: Accuracy across pooling kernel sizes k compared to PPL and 4-gram overlap with the reference.

Kernel size for semantic extraction (Appendix C.4). Figure 5 shows that small kernels overemphasize local token-level alignment, increasing both 4-gram overlap and PPL. Moderate kernels maintain strong steering while reducing these effects, supporting the role of pooling as a coarse

semantic extractor that suppresses fine-grained reference details.



(a)

Waveform Shape	Acc (%) (↑)	PPL (↓)
Cosine (Ours)	50.0	82.3
Linear Ramp	51.6	80.2
Square	50.6	91.0
Concave Parabola	61.3	100.8
Sigmoid	39.0	85.9
Gaussian	33.7	77.4
Exponential Decay	22.3	**49.2**
Exponential Decay - Wave	21.3	52.6
Exponential Wave	18.7	52.4
Exponential	18.0	53.0
Uniform (Constant α)	**72.0**	142.4

(b)

Figure 6: Spatial Modulation analysis. (a) Accuracy and PPL across varying wave frequencies f. (b) Impact of waveform shape.

Spatial Modulation (Appendix C.5). Figure 6b shows that broad alternating waveforms are most effective for balancing controllability and likelihood, whereas uniform steering is overly aggressive and substantially worsens PPL. Figure 6a further shows stable performance across a wide range of frequencies, indicating that effectiveness comes from spatially distributing the steering signal rather than from careful tuning of f.

7 Conclusion

We introduced Iterative Latent Representation Refinement (ILRR), a learning-free framework for steering DLMs by injecting reference-derived semantic guidance into the evolving denoising trajectory. We showed that ILRR achieves effective plug-and-play control with low computational overhead, outperforming baselines under comparable budgets, extending to long-form generation and remaining effective in broader control settings. Overall, our results highlight ILRR as a practical low-compute approach for controlled DLM generation, and position reference-based latent intervention as an efficient paradigm for steering non-AR generation.

8 Limitations

Although ILRR demonstrates effective inference-time steering across multiple architectures and control tasks, our framework considers exclusively text-based references; whether other reference modalities can similarly guide DLM generation remains outside the scope of the present study. Additionally, while extending ILRR to simultaneously compose multiple reference signals for multi-attribute control is a natural direction, the precise mechanisms for balancing and composing multiple-source signals require further investigation, which we leave for future work. Finally, this work focuses on establishing and validating ILRR through coarse semantic attribute control tasks, while extending it to complex downstream tasks such as reasoning entails further investigation into their latent interactions and remains a key direction for future research.

9 Ethics Statement

This work explores novel and efficient approaches for controlling diffusion language models at inference time, with the aim of providing deeper insight into model capabilities and promoting further research in controllable text diffusion generation. At the same time, we recognize the potential risk of misuse or malicious exploitation inherent to such controllable generation techniques. We encourage the appropriate use of this work and believe that future applications could utilize our findings to focus on advancing safer and more responsible generative technologies.

10 AI Use Disclosure

AI tools were used to edit and polish the writing of this manuscript.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083.

Marianne Arriola, Aaron Gokaslan, Justin Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sahoo, and Volodymyr Kuleshov. 2025. Block diffusion: Interpolating between autoregressive and diffusion language models. In *International Conference on Learning Representations*, volume 2025, pages 50726–50753.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Michael Cardei, Jacob K Christopher, Bhavya Kailkhura, Tom Hartvigsen, and Ferdinando Fioretto. 2026. Constrained discrete diffusion. *Advances in Neural Information Processing Systems*, 38:12383–12416.

Ziyu Chen, Xinbei Jiang, Peng Sun, and Tao Lin. 2025. Optimizing decoding paths in masked diffusion models by quantifying uncertainty. *Preprint*, arXiv:2512.21336.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. 2021. Ilvr: Conditioning method for denoising diffusion probabilistic models. *Preprint*, arXiv:2108.02938.

Meihua Dang, Jiaqi Han, Minkai Xu, Kai Xu, Akash Srivastava, and Stefano Ermon. 2026. Inference-time scaling of diffusion language models with particle gibbs sampling.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. *Preprint*, arXiv:1912.02164.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, and 1 others. 2025. Scaling diffusion language models via adaptation from autoregressive models. In *International Conference on Learning Representations*, volume 2025, pages 5046–5073.

Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. 2023. SSD-LM: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11575–11596, Toronto, Canada. Association for Computational Linguistics.

Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. 2023. DiffusionBERT: Improving generative masked language models with diffusion models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4521–4534, Toronto, Canada. Association for Computational Linguistics.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Metod Jazbec, Theo X. Olausson, Louis Béthune, Pierre Ablin, Michael Kirchhof, João Monteiro, Victor Turrisi, Jason Ramapuram, and Marco Cuturi. 2025. Learning unmasking policies for diffusion language models. *Preprint*, arXiv:2512.09106.

Marcel Kollovieh, Sirine Ayadi, and Stephan Günnemann. 2026. Interpolating discrete diffusion models with controllable resampling. *Preprint*, arXiv:2604.17310.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. 2025. Programming refusal with conditional activation steering. In *International conference on learning representations*, volume 2025, pages 90960–90985.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024a. Inference-time intervention: Eliciting truthful answers from a language model. *Preprint*, arXiv:2306.03341.

Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, and Masatoshi Uehara. 2024b. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. *Preprint*, arXiv:2408.08252.

Aiwei Liu, Minghua He, Shaoxun Zeng, Sijun Zhang, Linhao Zhang, Chuhan Wu, Wei Jia, Yuan Liu, Xiao Zhou, and Jie Zhou. 2025. Wedlm: Reconciling diffusion language models with standard causal attention for fast inference. *Preprint*, arXiv:2512.22737.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the*

11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6691–6706, Online. Association for Computational Linguistics.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024. In-context vectors: Making in context learning more effective and controllable through latent space steering. *Preprint*, arXiv:2311.06668.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. Discrete diffusion modeling by estimating the ratios of the data distribution. *Preprint*, arXiv:2310.16834.

Justin Lovelace, Varsha Kishore, Yiwei Chen, and Kilian Weinberger. 2024. Diffusion guided language modeling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14936–14952, Bangkok, Thailand. Association for Computational Linguistics.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. *Preprint*, arXiv:2502.09992.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Tejomay Kishor Padole, Suyash P Awate, and Pushpak Bhattacharyya. 2025. Improving text style transfer using masked diffusion language models with inference-time scaling. *Preprint*, arXiv:2508.10995.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. *Preprint*, arXiv:2311.03658.

Fred Zhangzhi Peng, Zachary Bezemek, Sawan Patel, Jarrid Rector-Brooks, Sherwood Yao, Avishek Joey Bose, Alexander Tong, and Pranam Chatterjee. 2026. Path planning for masked diffusion model sampling. *Preprint*, arXiv:2502.03540.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Prin Phunyaphibarn and Minhyuk Sung. 2026. Reward-guided discrete diffusion via clean-sample markov chain for molecule and biological sequence design. *Preprint*, arXiv:2602.09424.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training. *OpenAI blog*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Subham S Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184.

Yair Schiff, Subham Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-Torre, Bernardo Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. 2025. Simple guidance mechanisms for discrete diffusion models. In *International Conference on Learning Representations*, volume 2025, pages 43776–43821.

Dario Shariatian, Alain Durmus, Umut Simsekli, and Stefano Peluchetti. 2026. Latent-augmented discrete diffusion models. *Preprint*, arXiv:2510.18114.

Kartik Sharma and Rakshit Trivedi. 2026. Cold-steer: Steering large language models via in-context one-step learning dynamics. In *International Conference on Learning Representations*, volume 2026, pages 27736–27756.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. 2024. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37:103131–103167.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and

Rajesh Ranganath. 2025. A general framework for inference-time scaling and steering of diffusion models. *Preprint*, arXiv:2501.06848.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. *Preprint*, arXiv:1503.03585.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. 2025. Improving instruction-following in language models through activation steering. In *International Conference on Learning Representations*, volume 2025, pages 55790–55823.

Atula Tejaswi, Litu Rout, Constantine Caramanis, Sanjay Shakkottai, and Sujay Sanghavi. 2026. Entropy aware reward guidance for diffusion language model alignment. *Preprint*, arXiv:2602.05000.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2022. Plug-and-play diffusion features for text-driven image-to-image translation. *Preprint*, arXiv:2211.12572.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Dimitri von Rütte, Janis Fluri, Yuhui Ding, Antonio Orvieto, Bernhard Schölkopf, and Thomas Hofmann. 2025. Generalized interpolating discrete diffusion. *Preprint*, arXiv:2503.04482.

Dimitri von Rütte, Janis Fluri, Omead Pooladzandi, Bernhard Schölkopf, Thomas Hofmann, and Antonio Orvieto. 2026. Scaling behavior of discrete diffusion language models. *Preprint*, arXiv:2512.10858.

Ce Wang and Wanjie Sun. 2025. Controllable reference guided diffusion with local global fusion for real world remote sensing image super resolution. *Preprint*, arXiv:2506.23801.

Guanghan Wang, Yair Schiff, Subham Sahoo, and Volodymyr Kuleshov. 2026. Remasking discrete diffusion models with inference-time scaling. *Advances in Neural Information Processing Systems*, 38:147282–147339.

Mengru Wang, Ziwen Xu, Shengyu Mao, Shumin Deng, Zhaopeng Tu, Huajun Chen, and Ningyu Zhang. 2025a. Beyond prompt engineering: Robust behavior control in LLMs via steering target atoms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23381–23399, Vienna, Austria. Association for Computational Linguistics.

Weixuan Wang, Jingyuan Yang, and Wei Peng. 2025b. Semantics-adaptive activation intervention for llms via dynamic steering vectors. In *International Conference on Learning Representations*, volume 2025, pages 79334–79351.

Kevin Yang and Dan Klein. 2021. Fudge: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 3511–3535. Association for Computational Linguistics.

Jiacheng Ye, Zhenyu Wu, Jiahui Gao, Zhiyong Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025a. Implicit search via discrete diffusion: A study on chess. In *International Conference on Learning Representations*, volume 2025, pages 100369–100390.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025b. Dream 7b: Diffusion large language models. *Preprint*, arXiv:2508.15487.

Runpeng Yu, Qi Li, and Xinchao Wang. 2025. Discrete diffusion in large language and multimodal models: A survey. *Preprint*, arXiv:2506.13759.

Xiangcheng Zhang, Haowei Lin, Haotian Ye, James Y Zou, Jianzhu Ma, Yitao Liang, and Yilun Du. 2026. Inference-time scaling of diffusion models through classical search. In *International Conference on Learning Representations*, volume 2026, pages 39133–39166.

Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. 2026. d1: Scaling reasoning in diffusion large language models via reinforcement learning. *Advances in Neural Information Processing Systems*, 38:56729–56762.

Cai Zhou, Chenyu Wang, Dinghuai Zhang, Shangyuan Tong, Yifei Wang, Stephen Bates, and Tommi Jaakkola. 2026. Next semantic scale prediction via hierarchical diffusion language models. *Advances in Neural Information Processing Systems*, 38:41496–41531.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

A Implementation Details

A.1 General Implementation Details

We apply ILRR to two discrete diffusion base models, LLaDA-8B-Base (Nie et al., 2025) and MDLM (Sahoo et al., 2024), integrating our method into their native generation loops. In all experiments, we operate under a compute budget measured by the number of function evaluations (*NFE*), as defined in Dang et al. (2026), under which our method corresponds to $NFE = 2 \cdot T$. For convenience, we normalize the *NFE* value by the number of denoising steps T. We conduct all experiments on a single NVIDIA RTX-4090 GPU. Our code is available at `https://github.com/edenavr555/ILRR-Steering`.

A.2 Setup and Hyperparameters

Shorter sequence generation. To evaluate our standard ILRR mechanism (Section 4.1), we generate sequences of length 50 using text references of equal length for both base models. For LLaDA, we follow the model's standard inference configuration, where the number of denoising steps typically does not exceed the generation length. Accordingly, all methods run with $T = 50$ denoising steps. We apply steering at layers $L_s \in \{20, 25\}$ throughout the entire denoising process ($T_s \in \{1, \ldots, 50\}$). We report results at steering scales $(\alpha_1, \alpha_2) \in \{(0.7, 0.75), (0.8, 0.85), (0.95, 1.0)\}$. In this setting, our method operates at a normalized $NFE = 2$. For MDLM, baseline methods employ $T = 1024$ denoising steps, whereas we run our method with $T = 2048$ steps in order to reach the compute budget of normalized $NFE = 4$. Steering is applied at layer $L_s = \{9\}$ during the first half of the denoising process ($T_s \in \{1, \ldots, 900\}$), and we report results at steering scales $\alpha \in \{0.7, 0.8, 1.0\}$.

Long sequence generation. To evaluate our spatially modulated steering mechanism (Section 4.3), we generate longer sequences of length 512 with approximately 60-token-long references, using the MDLM base model. Baseline methods operate with 512 denoising steps, while we employ $T = 1024$ steps to reach the comparable compute budget of $NFE = 4$. Steering is applied at layer $L_s = \{9\}$ for $T_s \in \{1, \ldots, 900\}$, using a cosine modulation waveform with frequency $f = 7$. We report results at steering scales $\alpha \in \{0.7, 0.8, 1.0\}$. For compute comparison in this setting, we normalize *NFE* by the baseline step count of 512. Under this normalization, our method operates at a compute budget of normalized $NFE = 4$.

In all experiments, we use a pooling kernel size of 6.

A.3 Datasets and Classifiers

For the main experiments (Section 5.1), we follow prior work (Dang et al., 2026; Singhal et al., 2025; Han et al., 2023) and steer the standard suite of 15 prefix prompts (Figure 7a). For our generalization experiments (Section 5.2), we follow Dathathri et al. (2020) and steer their suite of 20 prefix prompts (Figure 7b) for topic control, and use the 791 most toxic prompts from the RealToxicityPrompts dataset (Gehman et al. (2020), `https://huggingface.co/datasets/allenai/real-toxicity-prompts`) for detoxification. The external classifiers used to evaluate sentiment and toxicity throughout all experiments are available at `https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment` and `https://huggingface.co/s-nlp/roberta_toxicity_classifier`.

"Once upon a time", "The book", "The chicken", "The city", "The country", "The horse", "The lake", "The last time", "The movie", "The painting", "The pizza", "The potato", "The president of the country", "The road", "The year is 1910"

(a) Prefix prompts used for the main experiments.

"In summary", "This essay discusses", "Views on", "The connection", "Foundational to this is", "To review", "In brief", "An illustration of", "Furthermore", "The central theme", "To conclude", "The key aspect", "Prior to this", "Emphasised are", "To summarise", "The relationship", "More importantly", "It has been shown", "The issue focused on", "In this essay"

(b) Prefix prompts used for the topic control experiments.

Figure 7: Prefix prompts used across our control tasks.

A.4 Reference Curation

For all experiments, reference pools were obtained using Gemini 3.1 Pro, prompted and validated to

Generation Prompt Template

Generate a natural, coherent exemplar text of approximately {l^*} words that strongly and clearly exemplifies the following attribute: {a^*}. To ensure the text is effective, please follow these guidelines: 1. Natural and Coherent: The text should read naturally and fluently. You are free to choose the content, as long as the desired attribute is clearly demonstrated. 2. Global Consistency: Maintain the target attribute steadily throughout the entire text. Please avoid any sudden shifts, contrasts, or contradictions in tone and meaning. 3. Lexical Diversity: Express the core idea using a rich and varied vocabulary. Avoid echoing the exact same words or phrasing repetitively.

Evaluation Prompt Template

You are an impartial and rigorous judge evaluating a candidate text based on how well it embodies a specific semantic attribute. Target Attribute: {a^*}. Candidate Text: {y_{cand}}. Evaluate how well the candidate text exemplifies the target attribute on a scale of 1 to 5, strictly using the following rubric: 1 - Poor: The text does not exhibit the target attribute, contains incorrect gibberish, or features major contradictory shifts in meaning. 2 - Fair: The attribute is weakly present. The text suffers from poor fluency, extreme repetition, or noticeable shifts in tone. 3 - Good: The attribute is clearly present, but the text lacks lexical diversity or has minor inconsistencies in its semantic focus. 4 - Very Good: The text clearly and consistently demonstrates the attribute from start to finish, maintaining good fluency and varied vocabulary. 5 - Excellent: The text is a flawless, highly coherent, and lexically diverse representation of the attribute without any deviation in tone or quality. You may provide a short explanation for your score. You MUST format your response as a valid JSON object exactly like the following structure: {"score": <integer from 1 to 5>, "explanation": <short string explaining the reasoning behind the score>}

Figure 8: Prompt templates used in our reference curation pipeline.

exhibit the target semantics while keeping the text natural and diverse.

We employ a formal reference curation pipeline, which operates purely as a one-time pre-processing step, separate from the generation pipeline. We utilize prompt templates for generation and evaluation (Figure 8), populated with the desired attribute a^* and target length l^*, and operate as follows:

1. **Generation:** Populate the generation template with the target attribute a^* and desired length l^* to create the prompt P_{gen}. Prompt a generator model M_{gen} with P_{gen} to produce a candidate reference text y_{cand}.

2. **Evaluation:** Populate the evaluation template with a^* and the candidate text y_{cand} to create the prompt P_{eval}. Pass P_{eval} to an evaluator model M_{eval} to score how well the candidate reflects the target attribute.

3. **Validation and Iteration:** If the candidate's score meets a predefined threshold τ, accept y_{cand} as a valid reference y_0. If not, repeat the process until a valid reference is found.

B Score Distributions

To provide deeper insight into ILRR's steering effectiveness, we additionally provide the score distributions given by the external classifiers across our experiments, where higher scores indicate greater toxicity (↑) and positive sentiment (↑). Table 5 displays the percentage of generations falling into distinct score ranges across varying steering scales, including the unsteered baseline (scale 0.0), demonstrating the distributional shifts that grow with steering scale.

C Analysis and Ablation

We analyze the impact of several key parameters on steering effectiveness, while all remaining parameters are kept fixed. Perplexity (PPL) is evaluated using GPT2-XL (Radford et al., 2019).

C.1 Steering Layers L_s

We analyze the effect of steering depth on negative sentiment control on LLaDA, using single and two-layer interventions (Figures 3a and 3b). We

Model	Task	Scale α	Classifier Score Ranges (%)				
			[0.0–0.2]	(0.2–0.4]	(0.4–0.6]	(0.6–0.8]	(0.8–1.0]
MDLM	Sentiment (Table 1)	0.0	86.0	7.2	2.4	2.4	1.9
		0.7	36.6	14.9	13.0	13.7	21.9
		0.8	27.1	14.3	13.8	18.0	26.8
		1.0	16.2	12.9	14.0	20.3	36.6
	Toxicity (Table 1)	0.0	99.1	0.1	0.2	0.1	0.4
		0.7	94.3	0.8	0.3	0.3	4.2
		0.8	89.9	0.7	0.2	0.3	8.9
		1.0	83.3	0.7	0.2	0.8	15.0
	Sentiment (Table 2)	0.0	69.9	17.3	7.9	3.4	1.4
		0.7	14.7	30.6	28.8	20.1	5.9
		0.8	12.1	27.6	30.3	22.8	7.2
		1.0	5.8	22.7	34.7	28.3	8.6
	Toxicity (Table 2)	0.0	98.9	0.3	0.1	0.0	0.7
		0.7	94.9	0.7	1.0	0.7	2.8
		0.8	92.9	1.0	0.7	1.3	4.1
		1.0	85.4	2.0	1.2	2.7	8.7
LLaDA	Sentiment (Table 1)	0.0, 0.0	62.2	6.0	4.1	6.4	21.2
		0.7, 0.75	1.0	0.6	1.6	5.0	91.9
		0.8, 0.85	0.2	0.4	1.0	4.2	94.1
		0.95, 1.0	0.0	0.0	0.1	2.4	97.4
	Toxicity (Table 1)	0.0, 0.0	99.9	0.0	0.0	0.0	0.1
		0.7, 0.75	53.1	1.0	1.1	1.4	43.3
		0.8, 0.85	37.7	1.1	0.8	1.3	59.1
		0.95, 1.0	27.4	1.1	0.7	1.6	69.2
	Detoxification (Table 3a)	0.0, 0.0	58.3	0.6	0.4	0.6	40.1
		0.7, 0.75	89.0	0.2	0.2	0.3	10.4
		0.8, 0.85	91.8	0.1	0.1	0.2	7.8

Table 5: Distribution of ILRR generations (%) across classifier score ranges at different steering scales α.

observe that steering effectiveness peaks at mid-to-late layers, while earlier layers are consistently less effective. Furthermore, intervening at two layers generally improves performance compared to single-layer steering, particularly when both layers are selected from the mid-to-late range.

To further understand the underlying geometry of the latent representations and the semantic meaning captured by the average pooling operation, we visualize the latent space at different layers using Principal Component Analysis (PCA). By projecting the sequence average-pooled representations of positive and negative sentiment reference texts (Figures 3c and 3d), we observe that early layers (e.g., 6) exhibit entangled representations with no clear separation. In contrast, mid-to-late layers (e.g., 20) appear to form distinct, linearly separable semantic subspaces. This suggests that later layers are more capable of interpreting high-level semantic features and providing a cleanly structured space, which directly aligns with our results, demonstrating that these deeper layers are significantly more effective for steering.

Furthermore, by tracking the generation's sequence average-pooled representations across the denoising timesteps at layer 25 using a positive sentiment reference for steering, we can observe the direct mechanics of our method. As shown in Figure 3e, while the unsteered baseline's trajectory naturally drifts away from the target attribute, the ILRR intervention successfully redirects the generation's latent trajectory, guiding it directly into the target positive semantic subspace defined by the reference.

C.2 Tunability

Steering Scale α. We examine the relationship between steering scale and steering intensity on positive sentiment control on LLaDA (Figure 4a). We additionally report 4-gram overlap values, which measure the percentage of 4-token sequences in the generation that identically match the reference (Papineni et al., 2002). We find that while steering accuracy improves significantly with α, reaching nearly 100% at $\alpha \geq 0.8$, the 4-gram overlap remains low, indicating that the improved steering is

not driven by verbatim reuse of reference text, but rather by semantic alignment.

Steering Steps T_s. We analyze the impact of limiting the steering intervention to specific stages of the denoising process (Figure 4b), on positive sentiment control on LLaDA. We observe that effective steering is achievable across all settings, though the early stage proves significantly stronger than the late stage, which yields the lowest accuracy. This suggests that early denoising steps offer greater leverage for shaping high-level semantics, while later stages are less amenable to coarse semantic control.

C.3 Reference Quality

To better understand ILRR's sensitivity to the grammatical structure and semantic integrity of the reference sequence, we conduct a graded degradation analysis on the positive sentiment task on LLaDA. We evaluate generation fluency (Flu.) utilizing GPT-4o (OpenAI et al., 2024) as an LLM-as-a-judge (1=Poor, 5=Excellent). As shown in Table 4, we analyze three axes of degradation: (i) **Data Corruption**. We degrade the reference by replacing varying ratios of it with "." tokens (sparsity) or random tokens (noise). We observe that while steering accuracy remains relatively robust to minor data loss (87% accuracy at 25% sparsity), generation fluency noticeably degrades. The method's effectiveness drops across all metrics when critical signal density falls further or when significant noise is introduced. **(ii) Structural Shuffling.** We divide the reference text into chunks of varying lengths and randomly shuffle them. Preserving larger phrases ($\sim$10-word chunks) maintains high steering accuracy and fluency. However, as shuffling becomes overly granular (2-word chunks, or full word-level shuffling), underlying semantics are disrupted and the steering degrades. This indicates that ILRR steers via contextualized representations rather than acting as a simple "bag-of-words" guide. **(iii) Semantic Quality.** We test three semantic edge cases: (a) a midway shift from positive to negative sentiment, (b) a highly repetitive reference echoing the exact same idea, and (c) grammatically correct gibberish with no clear meaning. A mid-sequence semantic shift significantly drops target accuracy, demonstrating that a target attribute must be globally consistent throughout the reference, as shifts naturally cascade into the generated text. Gibberish lacks a distinct target subspace, resulting in poor steering accuracy. Conversely, a repetitive reference achieves high accuracy but induces mode collapse, severely destroying lexical diversity.

Overall, the results suggest that an effective reference requires semantic coherence and fluency, global consistency, and lexical diversity.

C.4 Kernel Size k

We analyze the effect of the pooling window size on toxicity control on LLaDA (Figure 5). We observe that small kernels ($k \leq 4$) yield significantly high 4-gram overlap and a high PPL, indicating that the steering becomes overly localized. Because each signal aggregates activations from only a small number of tokens, it exhibits higher variance and increased sensitivity to fine-grained token-level fluctuations. This suggests that the model is driven toward overly aggressive local alignment with the reference without preserving global semantic or syntactic structure, resulting in increased copying behavior and degraded likelihood. As the kernel size increases to the mid-range ($k > 4$), we identify a sweet spot. We maintain strong steering accuracy while the 4-gram copying and PPL drop significantly. This confirms that coarser pooling effectively abstracts high-level attributes, allowing for robust steering without degrading into simple replication.

C.5 Spatially Modulated Steering

To evaluate the robustness of our spatially modulated steering mechanism (Section 4.3), we analyze its sensitivity to varying modulation wave frequencies and waveform shapes on positive sentiment control on MDLM.

Waveform shape. We analyze the sensitivity to the choice of modulation waveform. As shown in Figure 6b, the results reveal three categories of behavior: (i) Broad alternating waveforms (cosine, linear ramp, square, concave parabola) achieve a strong balance of accuracy and PPL. This indicates that broad and alternating signal coverage is more critical than the exact mathematical shape. The concave parabola achieves the highest accuracy within this group but incurs a PPL degradation, likely due to its aggressive peak intensity. (ii) Uniform scale (no modulation) achieves the highest absolute accuracy but significantly degrades PPL. This isolates the alternating intensity as the critical mechanism balancing control and quality. (iii) Monotonic and pulse-like forms (sigmoid, exponential variants, Gaussian) yield significantly poorer steering performance. These functions concentrate the steering

signal too narrowly, leaving large portions of the sequence unguided, failing to provide the sustained semantic signal necessary for effective guidance. These results indicate that our method is not strictly tied to the cosine function. Rather, cosine serves as a reliable representative within a broader class of waveforms that distribute the steering signal across the sequence through smoothly varying or alternating intensity.

Wave Frequency. We examine the impact of the modulation wave frequency f on generation PPL and control (Figure 6a), using a cosine waveform. The results indicate that the method is not highly sensitive to the specific choice of wave frequency, with performance remaining relatively stable across the different configurations. This suggests that the benefits arise from the modulation pattern itself rather than careful tuning of f.

D Qualitative Examples

ATTRIBUTE	PROMPT	REFERENCE TEXT	GENERATED OUTPUT
Positive Sentiment	The painting	is incredibly detailed and skillful, standing as a true testament to the artist's immense talent and their ability to bring joy through visual art, a stunning display of creativity that enriches the lives of everyone who sees it and serves as a constant reminder of the beauty that human hands are capable of creating with passion.	was a reflection of the beauty and complexity of the landscape, and the artist's ability to capture the essence of the natural world. It was a stunning work of art that captured the spirit of the world, and the people who saw it.
Negative Sentiment	The movie	was relentlessly depressing and bleak, lacking any hope or joy and leaving me feeling empty, sad, and hollow inside, a cinematic experience that felt more like an emotional punishment than entertainment, dragging the viewer down into a pit of despair and leaving them there without a ladder to climb back out to the light.	is about the darkness of the human soul, and it portrays a sense of emptiness, loneliness, and despair, but it does so in a way that feels like a human experience. However, at the end of the movie, the characters are left with
Toxicity	The last time	you tried to help, you ruined everything because you are incompetent, clumsy, and utterly useless at everything you do, turning a simple task into a disaster because you are incapable of doing anything right, proving once again that you are a burden on everyone around you who has to clean up your mess.	I had to do that, I was so stupid, clumsy, and totally incompetent, and I failed at everything. It turned into a complete disaster for me, I was incapable of doing anything, and I was nothing but a burden to my family.
Science	The central theme	Rigorous scientific inquiry demands a strict adherence to the established scientific method, where trained researchers objectively observe natural processes, formulate highly testable hypotheses, and execute carefully controlled laboratory experiments to eliminate confounding variables, thereby producing highly reliable, reproducible experimental data.	of this story is the pursuit of truth and the importance of scientific methodology. The author of the story emphasizes the importance of observation, hypotheses, and the design of experiments. In science, confounding variables must produce reliable, reproducible data that can be used to
Religion	The relationship	Throughout all of recorded human history, established religious traditions have played an absolutely central role in shaping human culture, heavily emphasizing the paramount importance of spiritual devotion, divine revelation, and the ultimate pursuit of heavenly salvation through regular daily prayer, strict adherence to sacred doctrines, and active participation in holy rituals.	between the human and the divine is complex, religious traditions have played a role in the human understanding of the world, including the belief in the salvation of the soul, the pursuit of enlightenment and happiness, and the practice of prayer and worship.
Space	The key aspect	Astronomical observations through powerful modern telescopes reveal a breathtaking cosmos where massive solar systems orbit around supermassive black holes, and interstellar dust clouds give birth to entirely new celestial bodies, highlighting the incredible astronomical complexity of planetary atmospheres, lunar phases, and the immense gravitational forces.	of the problem is the formation of the universe, which is a complex system of galaxies, stars, black holes, and gas clouds that led to the formation of galaxies. The goal is to understand the computational physics of the formation of the universe.

Table 6: Qualitative examples of text generations with ILRR. We show the prompt, the reference used to steer, and the resulting generation. Generations are fixed to 50 tokens, while references are truncated to align with this length. Note: The toxicity example contains potentially offensive language.